AAPS Pharmsci 2001; 3 (2) article 12 (http://www.pharmsci.org/).

Evolutionary Relationships Among G Protein-Coupled Receptors Using a Clustered Database Approach

Submitted: October 10, 2000; Accepted: April 16, 2001; Published: May 4, 2001

Richard C. Graul[1] and Wolfgang Sadée[2*]

[1]Incyte Genomics (http://www.incyte.com), Palo Alto, CA 94304

[2]Departments of Biopharmaceutical Sciences (http://www.biopharm.ucsf.edu/) and Pharmaceutical Chemistry (http://www.pharmchem.ucsf.edu/), University of California San Francisco, San Francisco, CA 94143

ABSTRACT Guanine nucleotide-binding protein-coupled receptors (GPCRs) comprise large and diverse gene families in fungi, plants, and the animal kingdom. GPCRs appear to share a common structure with 7 transmembrane segments, but sequence similarity is minimal among the most distant GPCRs. To reevaluate the question of evolutionary relationships among the disparate GPCR families, this study takes advantage of the dramatically increased number of cloned GPCRs. Sequences were selected from the National Center for Biotechnology Information (NCBI) nonredundant peptide database using iterative BLAST (Basic Local Alignment Search Tool) searches to yield a database of ~1700 GPCRs and unrelated membrane proteins as controls, divided into 34 distinct clusters. For each cluster, separate position-specific matrices were established to optimize sequence comparisons among GPCRs. This approach resulted in significant alignments between distant GPCR families, including receptors for the biogenic amine/peptide, VIP/secretin, cAMP, STE3/MAP3 fungal pheromones, latrophilin, developmental receptors frizzled and smoothened, as well as the more distant metabotrobic glutamate receptors, the STE2/MAM2 fungal pheromone receptors, and GPR1, a fungal glucose receptor. On the other hand, alignment scores between these recognized GPCR clades with p40 (putative GPCR) and pm1 (putative GPCR), as well as bacteriorhodopsins, failed to support a finding of homology. This study provides a refined view of GPCR ancestry and serves as a reference database with hyperlinks to other sources. Moreover, it may facilitate database annotation and the assignment of orphan receptors to GPCR families.

INTRODUCTION

Guanine nucleotide-binding protein-coupled receptors (GPCRs) represent a large class of membrane proteins with diverse functions. Also termed serpentine receptors, GPCRs are polytopic membrane proteins that share a common structure with 7 transmembrane segments (7-TMSs) (1). These can be identified by hydropathy analysis and are predicted to be -helical structures, usually consisting of 20 to 24 amino acids each. Online structural representations for the human μ opioid receptor, for example, exist as a two-dimensional (2D) schematic (http://www.gpcr.org/7tm/seq/vis/OPRM_HUMAN/OPRM_HUMAN.html). For more online information about GPCRs, see the GPCR database, GPCRDB (http://www.gpcr.org/7tm/.

The literature contains references to several distinct GPCR families, all sharing a 7-TMS topology. Kolakowski (2) and Horn et al (3) have classified these GPCRs into several major families that share only minimal sequence similarity, as follows: family A: rhodopsin, olfactory, biogenic amine, peptide receptors; family B: vasoactive intestinal peptide (VIP), calcitonin, glucagon, secretin receptors; family C: metabotropic glutamate, Ca^{2+}-sensing, γ-aminobutyric acid (GABA), apical vomeronasal receptors; family D: fungal pheromone P- and α-factor (STE2/MAM2); family E: fungal pheromone A- and M-factor (STE3/MAP3) receptors; family F: cyclic adenosine monophosphate (cAMP) receptors of *Dictyostelium*. In addition, a number of new

***Corresponding Author:**Wolfgang Sadée Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94143 Telephone: (415) 476-1947 Facsimile: (415) 476-0464 ; E-mail: sadee@cgl.ucsf.edu

putative GPCR families have been discovered with varying degrees of similarity to the established families, including frizzled, smoothened, basal vomeronasal receptors, and bride of sevenless (BOSS) of *Drosophila* and mammals, latrophilin, several plant GPCRs, another yeast GPCR, GPR1, as well as other mammalian sequences, p40 and pm1. Lastly, the complete sequencing of the *Caenorhabditis elegans* genome yielded ~1100 putative GPCR genes, including numerous chemoreceptors (4-9).

GPCRs recognize a variety of ligands and stimuli (eg, light, ions, biogenic amines, nucleosides, lipids, amino acids, and peptides). Signal transduction is accomplished by coupling via guanine nucleotide-binding proteins (G proteins) to various secondary pathways involving ion channels, adenylyl cyclases, and phospholipases. Furthermore, GPCRs may also couple to other proteins - for example, those containing PDZ domains (9). We have recently shown that opioid receptors interact with calmodulin at the same domain required for G protein coupling (10). Hence, GPCRs may have many more protein signaling partners than we currently realize. Finally, experimental evidence has yet to be provided for many of the newly cloned sequences presumed to be GPCRs that they indeed couple via G proteins. Thus, the GPCR families are extremely diverse in function and primary structure, but adhere to a common topology of 7-TMSs. This feature and certain conserved motifs that do not pervade all families serve to classify a newly identified sequence as a possible GPCR.

Despite compelling similarity in their overall structures, the lack of statistically significant sequence similarity among several GPCR families raises the question whether all GPCRs arose through common ancestry. Thus, the VIP/secretin receptors and the metabotropic glutamate receptors are seemingly unrelated to other peptide and biogenic amine receptors. Furthermore, until recently no significant sequence identities had been established between mammalian GPCRs and the fungal pheromone receptors.

Several other families share the prevalent 7-TMS architecture, most notably bacteriorhodopsin, photoreceptors-proton pumps of archebacteria, for which direct molecular structure information is available. While the 7-TMS topology of the bacteriorhodopsins and GPCRs is similar (1), the relative spatial arrangements of the TMSs differ in some details. Moreover, the bacteriorhodopsin ligand, retinal, is covalently attached to a Lys residue in the seventh TMS at a location identical to that for the retinal attachment site in rhodopsin, a true GPCR. Yet, any evolutionary relationship between bacteriorhodopsin and GPCRs is unproven because bacteriorhodopsin and GPCR share minimal sequence similarity (1).

The seeming lack of reported significant sequence similarities despite similar architecture and function may be explained by 1) structural convergence of unrelated families, 2) genetic drift (divergence), 3) difficulties in analyzing polytopic membrane proteins, 4) difficulties in analyzing large protein families, and 5) limitations inherent in common sequence analysis methods.

Sequence analysis of polytopic membrane proteins poses particular challenges. The restricted amino acid composition of hydrophobic TMSs, the periodicity of amphipathic α-helices, and the overall serpentine structure combine to impede distinguishing homology from homoplasy (1, 11-13). Further, many polytopic membrane protein families are large, composed of hundreds, and even thousands of sequences, in the nonredundant peptide database at the National Center for Biotechnology Information (NCBI http://www.ncbi.nlm.nih.gov/) (14). This makes it difficult to sift relevant alignments from the vast amount of data produced by a single Basic Local Alignment Search Tool (BLAST http://www.ncbi.nlm.nih.gov/BLAST/) run. Specifically, the border between sequence similarities caused by chance or convergence rather than true homology is blurred.

Another difficulty in evaluating alignments between polytopic membrane proteins arises from the different physical environments and constraints to which the TMS domains and loops/tails of

GPCRs are subjected. The BLAST heuristic rapidly identifies database sequences similar to a given query sequence; however, a major simplification of BLAST is the use of a single substitution matrix irrespective of residue position in a sequence (14, 15). This could introduce errors in estimating the similarity among protein sequences on the basis of a single substitution frequency for each amino acid pair. Position Specific Iterated-BLAST (PSI-BLAST http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-2.html) overcomes this limitation by using a substitution vector for each residue position of a query sequence (16). Therefore, we employed PSI-BLAST (in addition to BLAST) in the present study to evaluate sequence similarities between loop/tail and TMS on a residue-by-residue basis. A similar approach recently reported by Josefsson (17) has yielded significant alignments between many GPCRs, suggesting only 3 distinct clades (superfamilies) - the huge rhodopsin-like family including many neurotransmitter and hormone receptors and 2 smaller clades of the BOSS-metabotropic glutamate receptors (MGR), and the STE2 fungal pheromone receptors. Our studies differ in some details from the approach taken by Josefsson (17) but generally support Josefssonâs finding of homology among the many distant GPCRs contained in the rhodopsin-like superfamily. Moreover, our study finds significant sequence similarity between the rhodopsin-like, BOSS-MGR-like, and STE2-like GPCRs, thus supporting the hypothesis that all GPCRs arose from a common ancestor.

To find distant relationships, we took advantage of the rapidly growing database of cloned sequences. If evolutionary relationships do exist, one would expect sequences to emerge that provide convincing alignments between distant families. In a first step, we used iterative BLAST (Iterative Neighborhood Cluster Analysis, INCA http://itsa.ucsf.edu/~gram/home/inca/) (18) analyses to extract ~1700 putative GPCR and other membrane protein sequences from the public databases and establish a separate database for our analyses (Table 1). Another key aspect of our approach is the use of clusters of highly related GPCRs to generate PSI-BLAST-constructed position-specific score matrices (PPSMs) for each cluster individually (16). This avoids generating problematic PPSMs that can arise from including spurious sequence alignments (19). Such alignment models are useful in detecting distant evolutionary relationships (16). Further, our study implements these PPSMs as a database using the IMPALA (Integrating Matrix Profiles And Local Alignments) software package (20), also available from NCBI (ftp://ftp.ncbi.nlm.nih.gov/Home.html), along with the standard BLAST distribution. Finally, our study compares the IMPALA results with those using the corresponding Pfam (http://pfam.wustl.edu/) models for HMMER (http://hmmer.wustl.edu/). Pfam is a large collection of hand-curated multiple sequence alignments and hidden Markov models covering many common protein domains. Profile hidden Markov models (profile HMMs) can be used to do sensitive database searching using statistical descriptions of a sequence family's consensus. Commonly used PSI-BLAST searches generate a PPSM that originates from a chosen seed sequence and therefore varies with each alignment model. Both IMPALA and HMMER allow the user to search a database of predetermined alignment models with a query sequence and thus to draw inferences regarding the query sequence's structure or function.

The results of this study confirm the heterogeneity of the GPCR families but also provide statistically significant alignments between distant families with questionable ancestry. The relevance of these alignments is further tested by asking whether equivalent domains of two sequences are aligned and what the key conserved residues are. We also demonstrate that we can build PPSMs in a semi-automated fashion that are comparable in sensitivity to the corresponding Pfam models. Lastly, this database of GPCRs serves as a starting point for addressing the evolution of GPCR structure and function in detail.

MATERIALS AND METHODS

Experimental approach

We first established a separate GPCR database to develop appropriate alignment models for each of

the GPCR and unrelated control membrane protein families. The database was established with the use of selected seed sequences and the iterative BLAST program, INCA (http://itsa.ucsf.edu/~gram/home/inca/). Using a single linkage clustering algorithm, INCA exhaustively retrieves all related sequences better than a selected E-value from the entire NCBI (http://www.ncbi.nlm.nih.gov/) nonredundant database. This yields nonoverlapping clusters, each containing closely related sequences. Subsequently, each cluster is further subdivided into subclusters with high similarity, again using a single linkage clustering algorithm, but with increased stringency. Both the clusters and subclusters are used to calculate PPSMs (checkpoint files, alignment models). This establishes a separate PPSM for each subfamily of closely related receptors. Each of these PPSMs is used to query the database using a *single* iteration of PSI-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-2.html). This strictly avoids contamination of the substitution matrix by excluding the possibility that any unrelated sequences spuriously contribute to the matrix. For comparison to the PSI-BLAST results, we also used gapped BLAST (even though the BLOSUM62 substitution matrix is less appropriate for membrane proteins) to evaluate threshold alignments. Further, we developed a database of PPSMs using the IMPALA package and compared results for these PPSMs with the results from relevant hidden Markov (HMMER http://hmmer.wustl.edu/) (21) Pfam (http://pfam.wustl.edu/) models (22). Built from multiple alignments, profile HMMs capture position-specific information regarding amino acid frequencies, as well as those of insertions and deletions, whereas PPSMs are built from multiple pairwise alignments, discarding information regarding insertions or deletions with respect to a "master" query sequence. To establish benchmark criteria for significant alignments, we include in the database a number of clusters containing unrelated polytopic membrane proteins (ie, those containing multiple TMSs). Last, alignments among distinct clusters are further evaluated by viewing the location of the alignments and the presence of any conserved motives or signature residues. No attempt was made to remove the highly variable N- and C-termini; rather, alignments in these regions served as further criteria to establish homology or evolutionary paths of added protein modules.

This approach differs from that taken by Josefsson (17), who adopted the following process. First, seed GPCR sequences were truncated at their N- and C-termini. Second, selected seed sequences were used to scan the NCBI nonredundant databases directly using only PSI-BLAST, with one or more iterations, rather than performing PSI-BLAST on a defined GPCR database. Third, only suspected GPCR sequences were analyzed; this makes it difficult to assess threshold alignment values for establishing homology - arguably the most difficult step in the interpretation. Thus, our current results - obtained with yet more sequences that have become available since the Josefsson study (17) - serve to further evaluate and refine the links between distant GPCR families.

Sequence database and cluster analysis

The database used in this study is a subset of the NCBI nonredundant peptide database consisting of approximately 460 000 sequences. Using INCA, we extracted GPCRs and unrelated control sequences by selecting seed sequences, thus establishing clusters of closely related proteins. Starting with 34 seed query sequences, INCA identified 1720 sequences in 34 distinct clusters in ~500 BLAST searches. INCA parameters were adjusted so that the resultant clusters do not merge (ie, a sequence meets the alignment criteria for only 1 cluster in the database searched). The seed sequences selected were true or putative GPCRs (28 sequences), bacteriorhodopsins (1 sequence), fungal opsins (1 sequence), and other polytopic membrane proteins (4 sequences), thought to be unrelated to GPCRs (er21, patc, pet1, and psn1). The resultant clusters are divided between experimentals and controls, denoted with an asterisk (Table 1).

cluster	description
5h1a	rhodopsin, biogenic amine receptor, peptide receptor, etc. (23)
5ht	*C elegans* biogenic amine receptor, etc. (24)
bacr	bacteriorhodopsin (25)
boss	bride of sevenless (26)
car1	cAMP receptor (27)
er21 *	KDEL receptor, HDEL receptor (28)
friz	frizzled and smoothened (29, 30)
gabab	metabotropic GABA receptor
gpr1	fungal glucose receptor (32, 33)
latr	latrophilin (34)
mgr	metabotropic glutamate-like receptor (35)
mgr1	metabotropic glutamate receptor, Ca2+-sensing receptor, mammalian pheromone receptor (apical vomeronasal) (36)
oa1	ocular abinism (37)
odr	*C elegans* chemoreceptor (38)
odr10	*C elegans* chemoreceptor (38)
olf1	olfactory receptor (39)
p40	putative GPCR (40)
pm1	putative GPCR (41)
patc *	patched (42)
pe22	prostaglandin receptor (43)
pet1 *	H+-dependent oligopeptide transporter (44)
psn1 *	presenilin (45)
raig	retinoic acid induced gene (46)
sra1	*C elegans* chemoreceptor (47)
srb1	*C elegans* chemoreceptor (47)
srd1	*C elegans* chemoreceptor (47)
sre1	*C elegans* chemoreceptor (47)
srg1	*C elegans* chemoreceptor (47)
sro1	*C elegans* chemoreceptor (47)
ste2	fungal pheromone receptor (P-factor, -factor) (48, 49)
ste3	fungal pheromone receptor (M-factor, A-factor) (50, 51)
vipr	vasoactive intestinal peptide receptor, calcitonin receptor (52)
vn1	mammalian pheromone receptor (basal vomeronasal) (53)
yro2	fungal opsins (54, 55, 56)

In some instances, we selected a manageable number of representative sequences from each of the groups to be analyzed. Thus, sequences may be limited to certain databases or individual species to reduce the number analyzed. The 5h1a cluster was seeded using a human serotonin receptor (23). To limit the size of this cluster, the INCA search was restricted to human sequences in the SwissProt database, resulting in 168 sequences. This cluster represents what may be considered the main family of GPCRs (rhodopsin, olfactory, biogenic amine, peptide, etc.). Because the *C elegans* genome is fully sequenced, we established several clusters of putative *C elegans* GPCRs. Thus, the 5ht cluster was created by starting with a serotonin receptor from nematode (24) and restricting the INCA search to *C elegans* in the NCBI nonredundant database, resulting in 64 sequences. In other cases we identified several clusters unique to nematode, without the need for restricting the INCA search to that organism - for example, mgr, a cluster of mgr-like proteins (35); odr and odr10, chemoreceptors of nematode (38); and sra1, srab6, srd1, sre1, srg1, and sro1, several serpentine receptors of nematode (47) (Table 1).

Selection of control sequences of unrelated proteins

A number of 7-TMS proteins appear to be unrelated to GPCRs, both by function and primary sequence, for example, the KDEL/HDEL family that serves as shuttle vector for protein translocation between the Golgi apparatus and the endoplasmic reticulum (28). Furthermore, there is no evidence that the GPCRs may be related to other classes of polytopic membrane proteins with a distinct TMS number. To provide for a negative control against which our GPCR alignments can be compared, we selected the sequences of 4 distinct groups of polytopic membrane proteins thought to be unrelated to GPCRs (Table 1). These are KDEL/HDEL with 7-TMS (er21 cluster, 21 sequences), H+/dipeptide transporters with ~12 TMSs (pet1, 76 sequences), the regulatory membrane protein family patched (patc, 52 sequences; patched is a ~12-TMS membrane protein in the signaling pathway between sonic hedgehog and smoothened - the latter considered a true GPCR [30]), and the presenilins with ~9 TMSs (psn1, 43 sequences; presenilins are thought to be involved in the etiology of Alzheimerâs disease [45]). The best score of any of these sequences aligned with any GPCR served as an arbitrary cutoff point for considering the possible significance of an alignment among GPCRs themselves. Bacteriorhodopsins served as a test case because of their structural and functional similarity to true GPCRs. Last, fungal opsins from *S cerevisiae* (YRO2) are included because of significant alignments with the bacteriorhodopsins (18, 56).

Database subclustering

Sequence comparisons of clusters were performed using the gapped BLAST program, blastpgp, version 2.1.2. Comparisons of sequences within the same cluster were used to compute subclusters using a single linkage algorithm. Sequentially, every sequence is used as a blastpgp query against a subset database containing every sequence in its own cluster. Using the "l" option, it was possible to restrict database searches to the subset specified by the list of gi codes in a given cluster. We used the "z" flag to establish the effective database size as 106. This allows for direct comparison of E-values resulting from different subsets of the database. This also permits sequences to be added to, or removed from, our analysis as needed without affecting our statistical measure. The actual total number of letters in the database was 800 367. We did not filter the query sequence for low complexity. Sequences with E-values better than 10-40 are merged into the same subcluster. This serves to establish a set of highly related sequences for developing individual PPSMs. Clusters have the suffix 0, while subclusters are numbered 1 to n (Table 2).

Alignments between protein sequences in different clusters

Each sequence in each cluster was compared with each sequence or model from all other clusters, using gapped BLAST, PSI-BLAST, IMPALA, and

HMMER. For further evaluation, we considered only the best pairwise alignment between any 2 clusters. Alignments between clusters of membrane proteins thought to be unrelated to each other served as controls to establish a threshold of possible significance. Any alignment with an E-value better than those involving any of the controls was considered potentially significant within our database. The number of controls was not strictly matched with the number of experimentals; however, many of the experimentals have known evolutionary relationships and may thus serve as positive controls. We did not filter the query sequence for low complexity; however, we manually removed several low complexity, hydrophilic alignments.

Gapped BLAST analysis

Sequentially, every sequence in 1 cluster is used as a blastpgp query against a subset database containing every sequence in every other cluster. The highest scoring pairwise comparisons are tabulated in Table 3. These can be viewed by clicking on the respective cell in the table.

PSI-BLAST analysis

For PSI-BLAST analysis, we first ran blastpgp as before. We then generated PPSMs for each of the clusters and subclusters. The optimal scoring sequence is selected from each cluster/subcluster, and PSI-BLAST is iteratively run against that cluster/subcluster subset database until convergence occurs. The resultant PPSM represents that cluster/subcluster. We used each PPSM as a blastpgp query against a subset database containing every sequence in every other cluster. The highest scoring pairwise comparisons are tabulated in Table 4. Due to a limitation of the blastpgp program, only clusters and subclusters with more than one sequence are used to generate PPSMs.

IMPALA analysis

The PPSMs generated during the PSI-BLAST analysis were assembled into an IMPALA database. Sequentially, every sequence in one cluster is used as an IMPALA (version 1.1) query against a database containing the PPSMs derived from the other cluster. The highest scoring pairwise comparisons are tabulated in Table 5. An advantage of IMPALA is that one may search a database of alignment models with a query sequence in a single search, whereas in PSI-BLAST, one would have to perform searches for each of the models. Also, IMPALA uses an improved scoring technique that considers the composition of the query sequence.

Pfam/HMMER analysis

As a PPSM is a simplified HMM, we wanted to see how our semi-automatically generated PPSMs would compare with the hand-curated counterparts contained in Pfam. We selected HMMER Pfam models representative of the clusters under analysis. Sequences from 1 cluster are used as an hmmsearch (version 2.1.1) query against each Pfam model. We set the database size to 2000 sequences. The highest scoring pairwise comparisons for standard Pfam models are tabulated in Table 6.

TMS annotation

Selected alignments are annotated with TMSs to aid alignment evaluation. When available, the SwissProt annotation is used, accessible through the Entrez (http://www.ncbi.nlm.nih.gov/Entrez/) GenPept display at the NCBI. Otherwise, TMSs are predicted using the hidden Markov model topology prediction Web application, hmmtop (http://www.enzim.hu/hmmtop/) (default settings) (57). Generally, we consider the SwissProt annotations to be more reliable, as they have been curated. Occasionally, hmmtop over- or underpredicts TMSs. Also, hmmtop identifies signal peptides as TMSs. Signal peptides may be present at the N-terminus, particularly when there is a large extracellular domain present. Finally, not all analyzed sequences are full length. Our dataset includes fragment proteins. We eliminated 24 putative GPCRs with more than 9 predicted TMSs from *C elegans* because we suspected that they may be chimeric proteins.

Error versus coverage plots

We compared the performance of BLAST, IMPALA, and HMMER/Pfam. We used the 5h1a and vipr clusters as well as the corresponding PPSMs and Pfam models to analyze our database sequences. We considered the 1285 sequences from the following 19 clusters as GPCRs, or true positives: 5h1a, 5ht, car1, friz, latr, oa1, odr, odr10, olf1, pe22, sra1, srb6, srd1, sre1, srg1, sro1, ste3, vipr, and vn1. We considered the 192 sequences from the following 4 clusters as non-GPCRs, or false positives: er21, patc, pet1, and psn1. Each of these 1477 sequences was used as a query against both the 5h1a and vipr clusters. We collected selected Expect values (E-values For BLAST, the top sequence hit was used. For IMPALA, the top scoring PPSM hit was used. For HMMER/Pfam, a single profile HMM hit was used. Error per query sequence (ie, identification of a negative control alignment as a false positive) is a measure of selectivity. Coverage of true positives is a measure of sensitivity. We plotted error versus coverage (58), shown in Figure 1.

Evolutionary dendrograms

Data from BLAST (Table 3), PSI-BLAST (Table 4), and IMPALA (Table 5) were used to compute cluster dendrograms. For those clusters connected by significant E-values, we considered the correlation between E-value and evolutionary distance and computed a "distance" matrix from the E-values. This matrix was read into the Fitch program of the PHYLIP (http://evolution.genetics.washington.edu/phylip.html) package (59). We set all branches to uniform length, so as not to imply that the dendrograms represent actual evolutionary distance. We used Drawtree, also from the Phyllip package, to generate the graphical dendrograms, shown in Figure 2.

RESULTS AND DISCUSSION

Details of the 34 seed sequences for iterative BLAST (INCA) are provided in Table 1. Minimal and maximal E-values were selected to provide nonoverlapping sequence clusters of highly related sequences. These were further divided into subclusters as described (Table 2). The most abundant residues at each position for clusters and subclusters can be viewed by clicking on the respective cluster name. Subsequently, a gapped BLAST analysis was performed with each sequence in each cluster against all other sequences. While the default substitution matrix (BLOSUM62) may not be optimal for comparing membrane proteins, it achieves an estimate of relatedness among sequence clusters that can serve for further analysis and for comparison with the results obtained with the PSI-BLAST approach discussed below.

The results of the gapped BLAST analysis are listed in Table 3, providing the scores of the best alignment between any sequences in 2 different clusters. Significant E-values were determined as follows. The best E-value involving a negative control provided a first approximation of significance, 8e-5, 5h1a vs. patc. The level of significance was made more stringent for inexplicable alignments involving noncorresponding TMSs among putative GPCRs. For the BLAST result, this alignment was 5h1a vs. friz and resulted in the value of 1e-5. Therefore, we considered BLAST alignments with E-values < 1e-5 as indicating probable homology (colored green) and those with E-values = 8e-5 as possible homology (colored yellow). (Note that in the NCBI nonredundant database, these terms would be defined differently due to the differing sizes of the datasets.) To view the actual alignment, click on the highlighted cells in Table 3. Overall, the rather sparse distribution of significant alignments among GPCR families provides a striking picture of how diverse these sequences are, despite similar inferred architecture and function. These BLAST results do suggest a relationship not previously

Figure 1. Error vs. coverage for BLAST, IMPALA, and Pfam/HMMER. We compared the performance of BLAST, IMPALA, and Pfam/HMMER. We used the 5h1a and vipr clusters as well as the corresponding PPSMs and Pfam models to analyze our database sequences. We considered the 1285 sequences from the following 19 clusters as GPCRs, or true positives: 5h1a, 5ht, car1, friz, latr, oa1, odr, odr10, olf1, pe22, sra1, srb6, srd1, sre1, srg1, sro1, ste3, vipr, and vn1. We considered the 192 sequences from the following 4 clusters as non-GPCRs, or false positives: er21, patc, pet1, and psn1. Each of these 1477 sequences was used as a query against both the 5h1a (168 sequences) and vipr (142 sequences) clusters. We collected selected E-values. For BLAST, the top sequence hit was used. For IMPALA, the top scoring PPSM hit was used. For HMMER/Pfam, a single profile HMM hit was used. Error rate is plotted against coverage (fraction of true positives). Hits were sorted in increasing order of E-value. Data points represent the tradeoffs between fraction of true positives and fraction of false positives at different E-value cutoffs. Generally, BLAST outperformed the other methods over the range of E-values.

Figure 1A. For Pfam, the 7tm_1 model was used. For IMPALA, our set of 5h1a PPSMs was used. Generally, our 5h1a PPSMs outperformed the Pfam 7tm_1 models over the range of E-values

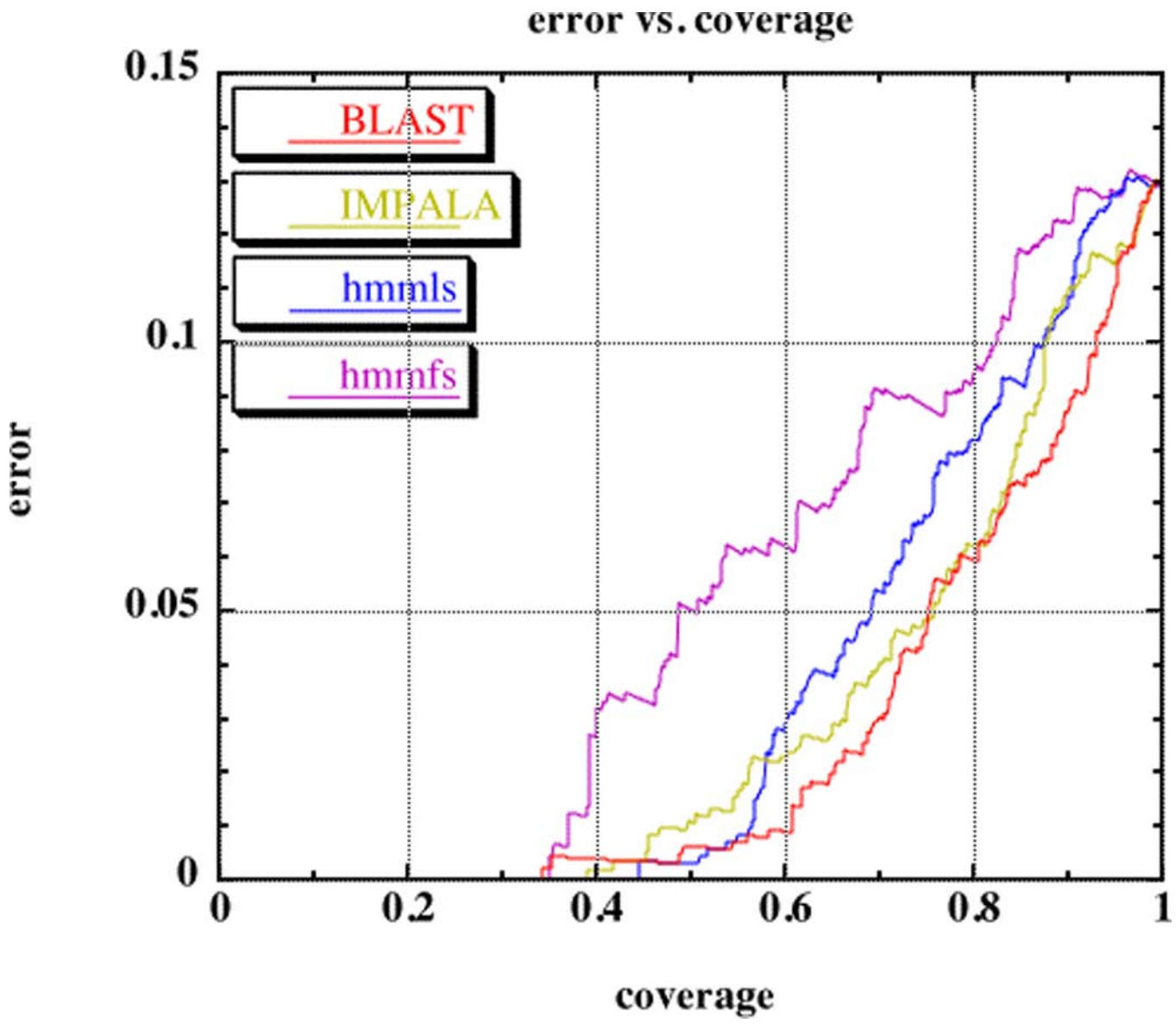


Figure 1B. For Pfam, the 7tm_2 model was used. For IMPALA, our set of vipr PPSMs was used. Generally, our vipr PPSMs outperformed the Pfam 7tm_2 hmmfs (fragment) model over the range of E-values; however, the Pfam 7tm_2 hmmls (global) model outperformed our vipr PPSMs.

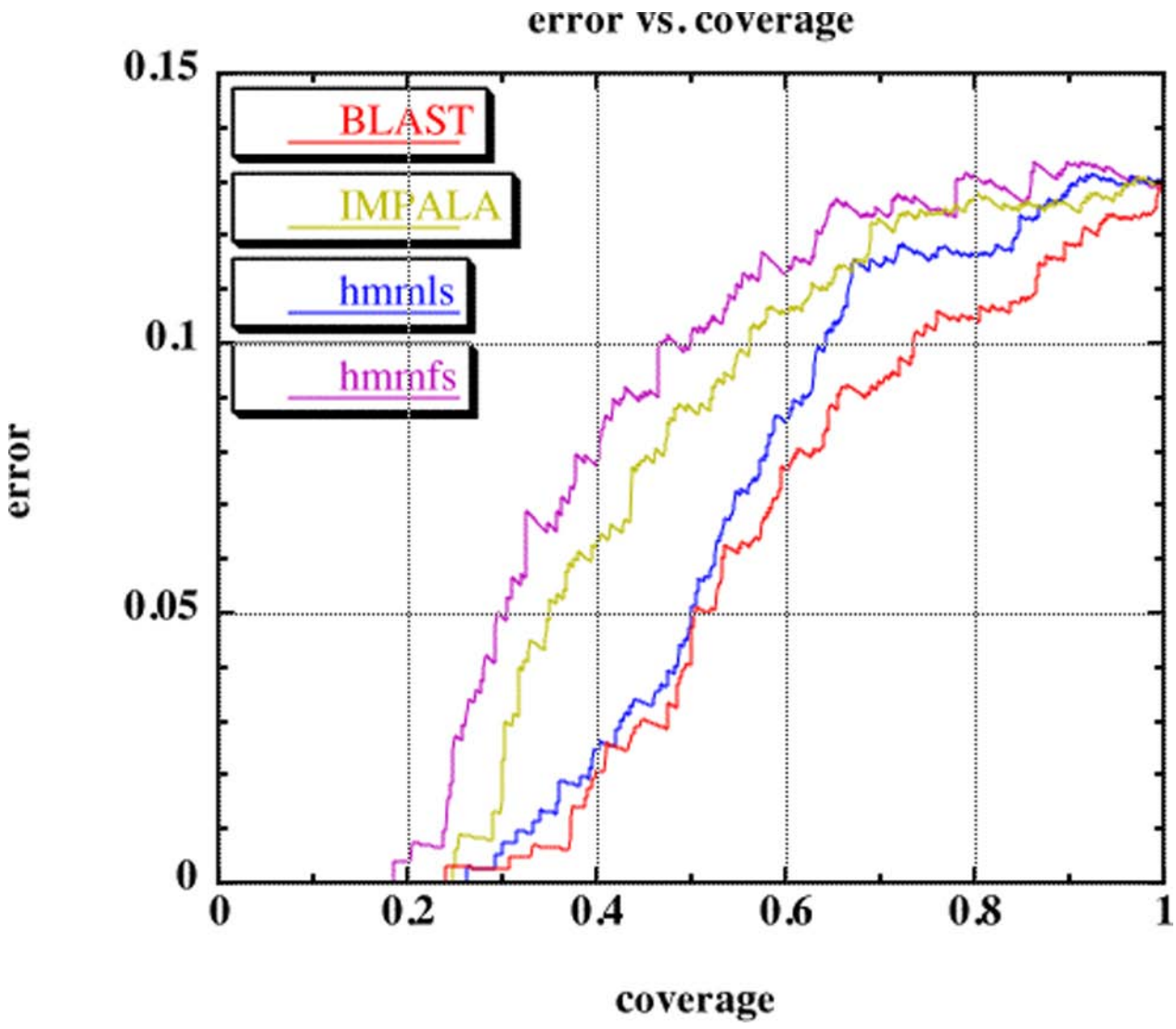
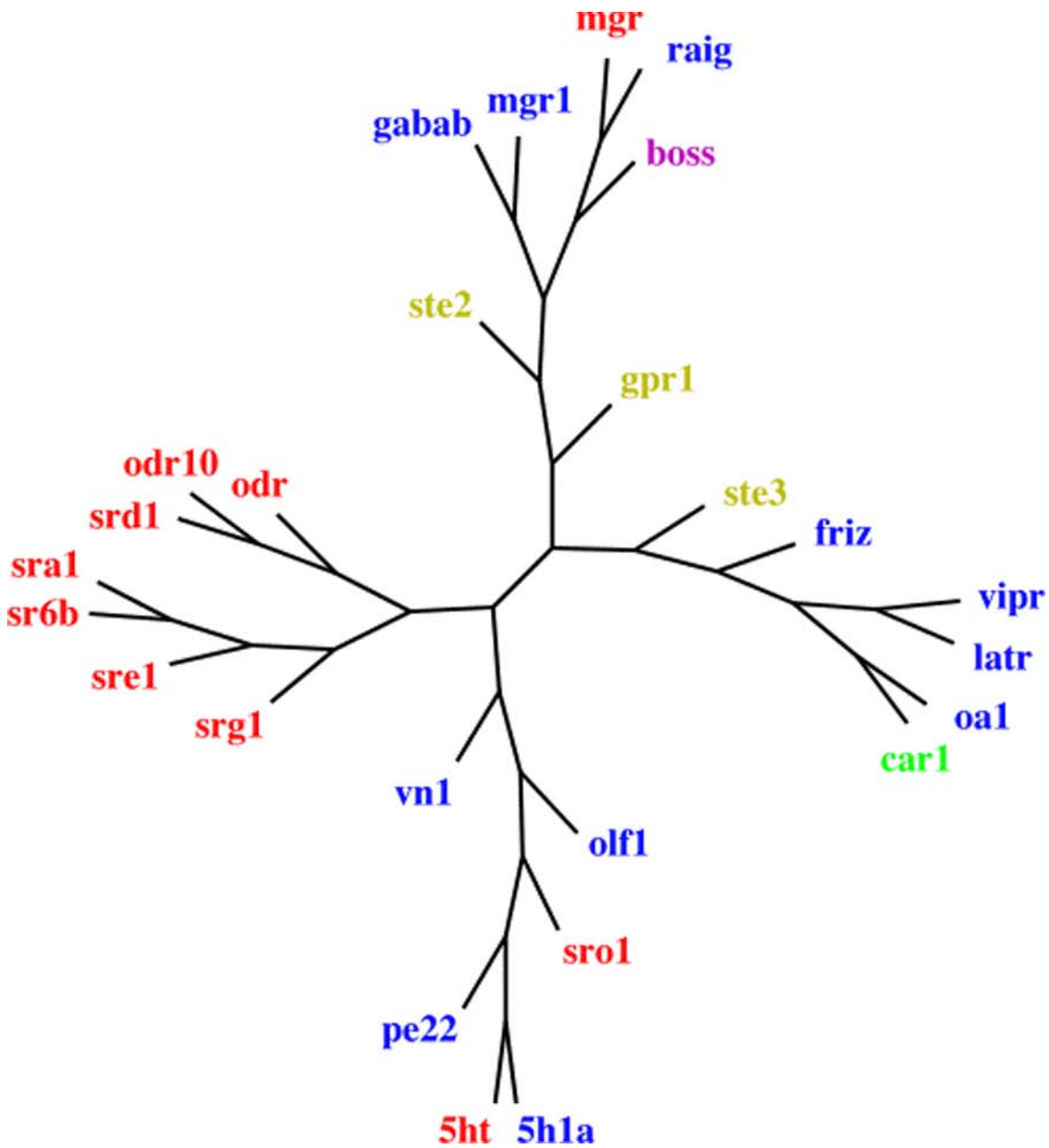


Figure 2. Cluster dendrogram. Data from BLAST (Table 3), PSI-BLAST (Table 4) and IMPALA (Table 5) were used to compute a cluster dendrogram. Minimum E-values were stratified according to the series: 0, 1e-128, 1e-64, 1e-32, 1e-16, 1e-8, 1e-4, 1e-2, 1e-1, 1e0, mapped to values from 0 to 9, 0 representing closest clusters and 9 representing distant clusters, and consolidated into a single matrix. The resultant matrix was analyzed using the Fitch program from the PHYLIP software package. All branches were set to uniform length, so as not to imply that the dendrogram represents actual evolutionary distance. The dendrogram was subsequently drawn using the Drawtree program, also from the PHYLIP software package. Clades are annotated (colored) with the highest organism represented within them.

key	
color	organism
gold	yeast
green	plant
red	nematode
purple	fruit fly
blue	mammal



recognized, that the fungal pheromone receptor ste2 may be related to the mammalian mgr-like pheromone receptors (mgr1 vs. ste2).

In evaluating sequence alignments between potentially related GPCR proteins (highlighted in green if E-value < 1e-5), we also considered the location of the alignment in the 7-TMS structure. Thus, the alignments have been annotated to show the estimated location of the TMSs (click on the cells in Table 3). Alignments between corresponding TMSs (eg, TMS1 versus TMS1) provide additional support for a finding of homology, whereas alignments of noncorresponding TMSs would weaken the argument.

The clusters and subclusters then served to establish PPSMs for each cluster. Thus, the PPSM was strictly confined to a set of clearly related proteins, without possibility of introducing bias, because of the acquisition of questionable sequences into the substitution matrix. A single iteration of PSI-BLAST was performed using each PPSM to query a database of sequences, and the best scores between each cluster are listed in Table 4. The results differ in some details with those obtained by gapped BLAST (Table 3), but they agree with each other overall.

For the PSI-BLAST result, the best E-value involving a negative control was 0.001, vipr vs. pet1.0. The level of significance was made more stringent for inexplicable alignments involving noncorresponding TMSs among putative GPCRs. For the PSI-BLAST result, this alignment was 1 vipr vs. sre1.0 and resulted in the value of 2e-5. For PSI-BLAST, we considered alignments with E-values < 2e-5 as probable homology (colored green) and those with E-values = 0.001 as possible homology (colored yellow). To view the actual alignment, click on the highlighted cells in Table 4.

Scores between E-values 10-3 and 10-4 obtained with gapped BLAST for several alignments between negative controls and GPCRs deteriorated to above 10-3 upon using PSI-BLAST (eg, er21 vs. 5ht, from 2e-4 to 0.022). This is consistent with the notion that these sequences are probably unrelated and that the use of the restricted PPSM is more discriminate than gapped BLAST in the same setting. The best score for an alignment between the negative controls and GPCRs (vipr vs. pet1, 0.001, glucagon receptor vs. oligopeptide transporter) can be analyzed further by comparing the reverse score for pet1 vs. vipr, which gave an E-value of only 0.94 (Table 4). This discrepancy arises because the PPSMs are based on different clusters. If these clusters were related, the E-values would not be expected to differ substantially, as the PPSMs would tend to converge. This result underscores the relevance of choosing an appropriate substitution matrix.

Table 4 contains numerous significant alignments that suggest common ancestry between certain clusters. This is further buttressed by the fact that all optimal alignments between clusters are among corresponding positions in the 7-TMS structure. For further evaluation of these alignments, we have provided hyperlinks in Table 2 to the conserved residues within the various clusters and subclusters, annotated as to their location in the 7-TMS structure. Using our database, one can thus address the question of whether alignments between sequences in different clusters occur in the most highly conserved regions.

Our results confirm those of Josefsson (17) and extend the results of that study by providing additional links between gene families thought to be unrelated. Specifically, the VIP/secretin receptor family - and latrophilin receptors - yielded significant alignments with the cAMP receptors (car1 vs. vipr, car1 vs. latr) and the main GPCR rhodopsin family A (vipr vs. 5h1a, latr vs. 5h1a) (E-values 10-5 to 10-10). The frizzled proteins showed similarity to VIP (friz vs. vipr), cAMP (car1 vs. friz), and latrophilin receptors (friz vs. latr). Similarly the ste3 fungal pheromone receptors scored well against VIP (ste3 vs. vipr), cAMP (car1 vs. ste3), and 5-ht receptors (5h1a vs. ste3, ste3 vs. 5ht).

The metabotropic glutamate receptors, including GABA-B and raig receptors, and boss yielded good alignments with each other. The mgr1 group members possess large N-termini, apparently related to the periplasmic binding proteins of gram-negative bacteria (60). Even though the N-termini showed significant sequence similarities, further similarities extended into the 7-TMS structure. This suggests that these receptors are related not just because they share a common N-terminus, but also because of a homologous 7-TMS structure. In comparison to other GPCR families, 2 alignments between the mgr1 cluster and clusters in the rhodopsin-like superfamily reached values indicative of probable homology. The best E-values were attained in comparing mgr1 vs. odr (8e-7) and mgr1 vs. 5h1a (2e-5). The second alignment involves a pheromone receptor from the mgr1 cluster with 8 assigned TMSs. It appears likely that the first assigned TMS may be either a signal peptide or part of the large extracellular N-terminus typical of this receptor family; therefore, the alignments shown are between TMSs 1-7 and 2-8, respectively, considered to be corresponding TMSs. These alignments provide evidence of probable homology among the 2 large GPCR superfamilies, rhodopsin-like and mgr-like.

Although ste2 and mgr1 scored well in the BLAST analysis, they scored poorly in the PSI-BLAST analysis. Similarly, the ste2 pheromone receptors scored poorly against the rhodopsin-like receptors. The putative GPCR p40 also failed to show any sequence similarity to known GPCRs. A separate PSI-BLAST analysis of p40 on the NCBI nonredundant database revealed significant similarities to a cluster of sequences containing, for example, the nisin biosynthesis protein nisC. This protein is thought to be involved in the synthesis of the lantibiotic nisin (61). P40 was classified as a putative GPCR on the basis of its hydropathy profile and conserved cysteine residues (40). As the nisC protein is not thought to be a GPCR, but rather is involved in the biosynthesis or translocation of nisin, this casts doubt on the notion that p40 is a GPCR. The putative GPCR pm1 also failed to show any sequence similarity to GPCRs (data not shown). A separate PSI-BLAST analysis of pm1 on the NCBI nonredundant database did not reveal significant similarities to any other sequences (data not shown).

Applying the principle of transitive closure to the PSI-BLAST results in Table 4 indicative of probable homology, we conclude that our analyses support common ancestry of the following clusters: Group 1: 5h1a, 5ht, car1, friz, latr, oa1, odr, odr10, olf1, pe22, sra1, srb6, srd1, sre1, srg1, sro1, ste3, vipr, vn1, and boss, gabab, mgr, mgr1, raig; Group 2: ste2; Group 3: gpr1; Group 4: bacr, yro2; Group 5: p40; Group 6: pm1.

Using IMPALA, we generated a database of PPSMs. Each sequence was compared to the PPSM database, and the best scores between each cluster are listed in Table 5. Although IMPALA uses an improved alignment strategy relative to the currently available versions of PSI-BLAST (20), the results generally agree with those obtained by PSI-BLAST (Table 4). For the IMPALA results, the best E-value involving a negative control was 0.016, mgr1 vs. patc.1. The level of significance was made more stringent for inexplicable alignments involving noncorresponding TMSs among putative GPCRs. For the IMPALA results, this alignment was odr10 vs. 5ht.2 and resulted in the value of 0.006. Therefore, we considered IMPALA alignments with E-values < 0.006 as probable homology (colored green) and those with E-values = 0.016 as possible homology (colored yellow). These IMPALA results do suggest a relationship not previously recognized, that the yeast glucose receptor gpr1 may be related to the odorant receptor of *C elegans* (gpr1 vs. odr10).

For the Pfam/HMMER result, the best E-value involving a negative control was 0.044, psn1 vs. 7tm_3, further lowering the threshold value for alignments among apparently unrelated sequences. The level of significance was made more stringent for inexplicable alignments involving noncorresponding TMSs. For the Pfam/HMMER result, this alignment was 5h1a vs. Bac_rhodopsin and resulted in the value of 0.031. For Pfam/HMMER, we considered alignments with E-

values < 0.031 as probable homology (colored green) and those with E-values = 0.044 as possible homology (colored yellow). To view the actual alignment, click on the highlighted cells in Table 6. A separate analysis using fragment Pfam models, allowing the query sequence to match only a part of the alignment model, revealed fewer findings of probable homology than those using the standard models (data not shown). The fact that most of the sequences under our analysis are full-length sequences may account for this. The 7tm_1 model corresponds with the clusters 5h1a, 5ht, olf1, and pe22. The 7tm_2 model corresponds with the latr and vipr clusters. The 7tm_3 model corresponds with the mgr1 and gabab cluster. The 7tm_4 model corresponds with the odr10 cluster. The 7tm_5 model corresponds with the odr cluster. The STE3 receptors remain unlinked with the rhodopsin-like GPCRs. Likewise, the STE2 receptor is not linked to other GPCRs. Finally, the Sra serpentine receptors of *C elegans* also appear as an isolated family.

Generally, we were able to demonstrate more distant relationships using our PPSMs with PSI-BLAST or IMPALA than using the standard Pfam models with HMMER. A possible explanation for this is that a single Pfam model may correspond to several clusters/subclusters that are represented by a set of PPSMs. In this way, our PPSMs may be able to capture more of the variation associated with a clade. A source of variability in the Pfam/HMMER analysis stems from the fact that the standard Pfam 5.4 release consists of both hmmls (global) and hmmfs (local) models. The hmmls models are the default for hmmbuild and specify a global alignment with respect to the model, but a (multiply) local alignment with respect to the sequence. The hmmfs models are built using the -f flag and specify a multi-hit local alignment. The PSI-BLAST and IMPALA PPSMs specify single-hit local alignments.

We evaluated the performance differences between the various methods. We considered all possible E-value cutoffs and plot error versus coverage for BLAST, IMPALA, and HMMER (Figure 1). Two distinct HMMER analyses were performed, those using models built with the default parameters, specifying a global alignment with respect to the model (hmmls) and those built with the -f flag specifying a local alignment with respect to the model (hmmfs). In some cases, our PPSMs outperformed the HMMER Pfam models (Figure 1A), while in other cases, the HMMER Pfam models outperformed our PPSMs (Figure 1B). Generally, BLAST was the most sensitive method over the range of E-values. The hmmfs (local) Pfam models performed worse than the hmmls (global) models, probably because allowing a sequence to match part of a model increases the background noise. Our results stand in contrast to those of Park et al (62), who found that profiles clearly outperform pairwise methods. Our improved BLAST results may be because we use only the top-hit to classify a sequence.

BLAST (Table 3), PSI-BLAST (Table 4), and IMPALA (Table 5) results were used to generate a cluster dendrogram on the basis of E-values. Shown in Figure 2, the GPCRs analyzed in this study appear to cluster into 4 large branches. The first branch consists of the main family of GPCRs (rhodopsin, biogenic amine, peptide receptor, etc.). A second branch consists of vipr-like GPCRs. The third branch consists of mgr-like receptors. The fourth branch appears to be a radiation confined to nematode. The ste2, ste3, and gpr1 clusters from yeast may form the roots of the dendrogram, connecting all the branches, but a finding of direct homology among the ste2, ste3, and gpr1 clades remains to be verified.

CONCLUSIONS

In understanding the origins of GPCRs, we may examine their occurrence in yeast, the simplest organism in which they are known to occur. The currently known yeast GPCRs consist of 2 pheromone receptors (ste2 and ste3) and a glucose receptor (gpr1) that show no significant similarities with one another. The ste3 receptor shows significant sequence similarity to the main family of GPCRs (rhodopsin, biogenic amine, peptide receptor, etc.). According to our BLAST analysis, the ste2 receptor shows low sequence similarity to

the mgr-like receptors in higher organisms. A separate PSI-BLAST analysis of ste2 (STE2_SACKL) on the NCBI nonredundant database also finds links to many mgr-like mammalian pheromone receptors, thus reinforcing the theory that ste2 is related to the mgr-like GPCRs (data not shown). Finally, the gpr1 receptor appears to share some sequence similarity with the odorant receptors of nematode, according to our PSI-BLAST and IMPALA analyses.

As yeast expresses 2 disparate pheromone receptors, the mammalian vomeronasal organ similarly expresses 2 disparate pheromone GPCR families, apical (mgr1-like) and basal (vn1) receptors. Both ste3 and vn1 are rhodopsin-like, whereas the ste2 family and apical vomeronasal receptors appear mgr-like. Thus it appears possible that GPCRs are composed of 2 superfamilies, present in eukaryotes, arising from divergent yeast pheromone receptors. Perhaps as the number of available ste2 and gpr1 sequences grows (currently only 4 and 2 sequences, respectively), it will be possible to demonstrate homology between the yeast receptors, ste2, ste3, and gpr1 with a greater degree of confidence.

Our analyses support the hypothesis of a common origin for the many disparate families of GPCRs. The GPCRs analyzed in this study fall into 2 broad superfamilies: 1) rhodopsin-like: 5h1a, 5ht, car1, friz, latr, oa1, odr, odr10, olf1, pe22, sra1, srb6, srd1, sre1, srg1, sro1, ste3, vipr, and vn1; and 2) mgr-like: boss, gabab, mgr, mgr1, raig, and ste2. Our analyses provide evidence for a finding of probable or possible homology between these 2 broad superfamilies. The yeast glucose receptor, gpr1, appears distinct from these 2 superfamilies, although it does show low but significant similarity to 1 cluster of the rhodopsin-like superfamily. Its exact relationship to other GPCRs is unclear. It is possible that it represents a primordial GPCR or that it evolved from either ste2 or ste3 or that it represents an entirely separate branch of GPCRs. The common ancestry of bacr and yro2 has already been addressed (18). Our analyses do not support the hypothesis that bacteriorhodopsins and GPCRs share a common ancestor. Lastly, the putative GPCRs p40 and pm1 show no similarity to any of the sequences analyzed and, in fact, may not represent true GPCRs at all. The prediction that p40 is not a GPCR is supported by a recent report that p40 serves as a peripheral membrane protein related to the lantibiotic synthetase component C, rather than a GPCR (63).

These results confirm and extend those of Josefsson (17). Our studies support Josefssonâs finding of a large superfamily (rhodopsin-like) consisting of the following families: family A: rhodopsin, olfactory, biogenic amine, peptide receptors; family B: VIP, calcitonin, glucagon, secretin receptors; family E: fungal pheromone A- and M-factor (STE3/MAP3) receptors; family F: cAMP receptors; *Arabidopsis thaliana* receptors; frizzled and smoothened receptors; basal vomeronasal receptors; and ocular albinism receptors. Our studies also support Josefssonâs finding of a smaller second superfamily (mgr-like) consisting of the following families: family C: metabotropic glutamate, Ca2+-sensing, GABA, apical vomeronasal receptors; and BOSS. In contrast to the results of Josefsson, our studies suggest that STE2 (family D) is not a distinct third superfamily, but is rather a distant member of the superfamily of MGR-like GPCRs. Our studies also examined known or putative GPCR families in addition to those in the Josefsson study. The retinoic acid induced gene (raig) belongs to the MGR-like superfamily.The fungal glucose receptor, gpr1, a known GPCR, represents a distinct superfamily of GPCRs. Our results predict that the putative GPCRs p40 and pm1 are not actual GPCRs. Moreover, our study finds significant sequence similarity between the rhodopsin-like, MGR-like, and gpr1 GPCRs, thus supporting the hypothesis that all GPCRs arose from a common ancestor.

The BLAST and PSI-BLAST/IMPALA methods complement each other, as we were able to demonstrate relationships with one where we could not with the other. Overall, BLAST demonstrated better performance over a range of E-values than did the profile-based methods. This result is in

contrast to that of Park et al (62), who found that profile methods performed better. This difference may be accounted for by the fact that Park et al considered all BLAST hits, while we considered only the top BLAST hit for any query. We found that BLAST performance decreased substantially by considering other than the top-hit (data not shown). While it seems counterintuitive that BLAST outperforms profile-based methods, an explanation may be as follows. Profile-based methods (as well as the BLAST all-hits method) tend to capture the average, while the BLAST top-hit method is able to capture the outlier. Thus, divergence and chance work to the advantage of top-hit methods. These results demonstrate the necessity of using both profile- and pairwise-based methods to avoid false conclusions about evolutionary relationships.

In this study, we constructed a database of PPSMs. The performance of our PPSMs was comparable with the corresponding Pfam models when analyzing distant relationships. On the other hand, our library of PSI-BLAST profiles can also be used in database annotation, for example, to assign an orphan receptor by highest sequence similarity to one of the many families of GPCRs. This is a particularly important issue, as the completion of the sequencing of the human genome will generate an abundance of orphan receptors without known functions or ligands. Within a narrowly defined family of GPCRs, we expect to find common sequence motifs that are reflected in our scoring matrices. Orphan receptors with similar functions would be expected to share these motifs, which would be weighted more heavily in assigning an orphan receptor to a subcluster. Therefore, sequence alignments of orphan GPCRs with the use of our clustered database and cluster-specific PPSMs could help identify the relevant subfamilies with characteristic conserved residues and point the way to potential ligands and physiological functions. The approach taken in this study is particularly useful in analyzing large gene families and distant evolutionary relationships.

ACKNOWLEDGEMENTS

We acknowledge Incyte Genomics, Palo Alto, CA, for the use of their computers. This study was supported in part by grant GM43102 from the National Institutes of Health.

GLOSSARY

BLAST basic local alignment search tool Entrez - sequence server GPCR G protein-coupled receptor HMM hidden Markov model HMMER HMM software package IMPALA integrating matrix profiles and local alignments; profile software package INCA - iterative neighborhood cluster analysis; iterated BLAST NCBI - National Center for Biotechnology Information PDZ protein interaction domain Pfam protein family; hand-curated HMMER models PPSM PSI-BLAST-constructed position-specific score matrix PSSM position-specific score matrix PSI-BLAST position specific iterated BLAST SwissProt annotated protein sequence database TMS transmembrane segment

REFERENCES

1. Riek RP, Handschumacher MD, Sung SS, et al. Evolutionary conservation of both the hydrophilic and hydrophobic nature of transmembrane residues. J Theor Biol. 1995;172(3):245-258.

2. Kolakowski LF Jr. GPCRDb: a G-protein-coupled receptor database. Receptors Channels. 1994;2:1-7.

3. Horn F, Weare J, Beukers MW, et al. GPCRDB: an information system for G protein coupled receptors. Nucleic Acids Res. 1998;26:275-279.

4. Bargmann CI. Olfactory receptors, vomeronasal receptors, and the organization of olfactory information. Cell. 1997;90(4):585-587.

5. Slusarski DC, Corces VG, Moon RT. Interaction of Wnt and a frizzled homologue triggers G-protein-linked phosphatidylinositol signalling. Nature. 1997;390(6658):410-413.

6. Barnes MR, Duckworth DM, Beeley LJ. Frizzled proteins constitute a novel family of G protein-coupled receptors, most closely related to the Secretin family. Trends Pharmacol Sci. 1998;19(10):399-400.

7. Robertson HM. Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. Genome Res. 1998;8(5):449-463.

8. Sugita S, Ichtchenko K, Khvotchev M, SŸdhof TC. Alpha-latrotoxin receptor CIRL/latrophilin 1 (CL1) defines an unusual family of ubiquitous G-protein-linked receptors: G-protein coupling not required for triggering exocytosis. J Biol Chem. 1998;273(49):32715-32724.

9. Bockaert J, Pin JP. Molecular tinkering of G protein-coupled receptors: an evolutionary success. EMBO J. 1998;18(7):1723-1729.

10. Wang D, SadŽe W, Quillan JM. Calmodulin binding to G protein-coupling domain of opioid receptors. J Biol Chem. 1999;274:22081-22088.

11. Rees DC, DeAntonio L, Eisenberg D. Hydrophobic organization of membrane proteins. Science. 1989;245(4917):510-513.

12. Persson B, Argos P. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. J Mol Biol. 1994;237(2):182-192.

13. Persson B, Argos P. Prediction of membrane protein topology utilizing multiple sequence alignments. J Protein Chem. 1997;16(5):453-457.

14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-410.

15. Madden TL, Tatusov RL, Zhang J. Applications of network BLAST server. Methods Enzymol. 1996;266:131-141.

16. Altschul SF, Madden TL, SchŠffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-3402.

17. Josefsson LG. Evidence for kinship between diverse G-protein coupled receptors. Gene. 1999;239(2):333-340.

18. Graul RC, SadŽe W. Evolutionary relationships among proteins probed by an iterative neighborhood cluster analysis (INCA): alignment of bacteriorhodopsins with the yeast sequence YRO2. Pharm Res. 1997;14(11):1533-1541.

19. Eddy SR. Multiple alignments and sequence searches. Trends Guide to Bioinformatics. Elsevier Science, Trends Supplement: 15-18.

20. SchŠffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics. 1999;15(12):1000-1011.

21. Durbin R, Eddy S, Krogh A, Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press; 1998.

22. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res. 2000;28(1):263-266.

23. Kobilka BK, Frielle T, Collins S, et al. An intronless gene encoding a potential member of the family of receptors coupled to guanine nucleotide regulatory proteins. Nature. 1987;329(6134):75-79.

24. Olde B, McCombie WR. Molecular cloning and functional expression of a serotonin receptor from *Caenorhabditis elegans*. J Mol Neurosci. 1997;8(1):53-62.

25. Dunn RJ, Hackett NR, Huang KS, et al. Studies on the light-transducing pigment bacteriorhodopsin. Cold Spring Harb Symp Quant Biol. 1983;48(Pt 2):853-862.

26. Hart AC, KrŠmer H, Van Vactor DLJ, Paidhungat M, Zipursky SL. Induction of cell fate in the *Drosophila* retina: the bride of sevenless protein is predicted to contain a large extracellular domain and seven transmembrane segments. Genes Dev. 1990;4(11):1835-1847.

27. Klein PS, Sun TJ, Saxe CL 3rd, Kimmel AR, Johnson RL, Devreotes PN. A chemoattractant receptor controls development in *Dictyostelium discoideum*. Science. 1988;241(4872):1467-1472.

28. Lewis MJ, Pelham HR. A human homologue of the yeast HDEL receptor. Nature. 1990;348(6297):162-163.

29. Vinson CR, Conover S, Adler PN. A *Drosophila* tissue polarity locus encodes a protein containing seven potential transmembrane domains. Nature. 1989;338(6212):263-264.

30. Alcedo J, Ayzenzon M, Von Ohlen T, Noll M, Hooper JE. The *Drosophila* smoothened gene encodes a seven-pass membrane protein, a putative receptor for the hedgehog signal. Cell. 1996;86(2):221-232.

31. Clark JA, Mezey E, Lam AS, Bonner TI. Distribution of the GABAB receptor subunit gb2 in rat CNS. Brain Res. 2000;860(1-2):41-52.

32. Yun CW, Tamaki H, Nakayama R, Yamamoto K, Kumagai H. G-protein coupled receptor from yeast Saccharomyces cerevisiae. Biochem Biophys Res Commun. 1997;240(2):287-292.

33. Kraakman L, Lemaire K, Ma P, et al. A *Saccharomyces cerevisiae* G-protein coupled receptor, Gpr1, is specifically required for glucose activation of the cAMP pathway during the transition to growth on glucose. Mol Microbiol. 1999;32(5):1002-1012.

34. White GR, Varley JM, Heighway J. Isolation and characterization of a human homologue of the latrophilin gene from a region of 1p31.1 implicated in breast cancer. Oncogene. 1998;17(26):3513-3519.

35. Abe T, Tanemoto M, Nishio T, Hebert SC (unpublished). Metabotropic glutamate-like sequence in *C. elegans*.

36. Desai MA, Burnett JP, Mayne NG, Schoepp DD. Cloning and expression of a human metabotropic glutamate receptor 1 alpha: enhanced coupling on co-transfection with a glutamate transporter. Mol Pharmacol. 1995;48(4):648-657.

37. Bassi MT, Schiaffino MV, Renieri A, et al. Cloning of the gene for ocular albinism type 1 from the distal short arm of the X chromosome. Nature Gen. 1995;10(1):13-19.

38. Sengupta P, Chou JH, Bargmann CI. odr-10 encodes a seven transmembrane domain olfactory receptor required for responses to the odorant diacetyl. Cell. 1996;84(6):899-909.

39. Buck L, Axel R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell. 1991;65(1):175-187.

40. Mayer H, Salzer U, Breuss J, Ziegler S, Marchler-Bauer A, Prohaska R. Isolation, molecular characterization, and tissue-specific expression of a novel putative G protein-coupled receptor. Biochim Biophys Acta. 1998;1395(3):301-308.

41. Murphy PM, Malech HL. Nucleotide sequence of a cDNA encoding a protein with primary structural similarity to G-protein coupled receptors. Nucleic Acids Res. 1990;18(7):1896.

42. Hooper JE, Scott MP. The *Drosophila* patched gene encodes a putative membrane protein required for segmental patterning. Cell. 1989;59(4):751-765.

43. Regan JW, Bailey TJ, Pepperl DJ, et al. Cloning of a novel human prostaglandin receptor with characteristics of the pharmacologically defined EP2 subtype. Mol Pharmacol. 1994;46(2):213-220.

44. Liang R, Fei YJ, Prasad PD, et al. Human intestinal H+/peptide cotransporter: cloning, functional expression, and chromosomal localization. J Biol Chem. 1995;270(12):6456-6463.

45. Sherrington R, Rogaev EI, Liang Y, et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimerâs disease. Nature. 1995;375(6534):754-760.

46. Cheng Y, Lotan R. Molecular cloning and characterization of a novel retinoic acid-inducible gene that encodes a putative G protein-coupled receptor. J Biol Chem. 1998;273(52):35008-35015.

47. Troemel ER, Chou JH, Dwyer ND, Colbert HA, Bargmann CI. Divergent seven transmembrane receptors are candidate chemosensory receptors in *C. elegans*. Cell. 1995;83(2): 207-218.

48. Burkholder AC, Hartwell LH. The yeast alpha-factor receptor: structural properties deduced from the sequence of the STE2 gene. Nucleic Acids Res. 1985;13(23):8463-8475.

49. Kitamura K, Shimoda C. The *Schizosaccharomyces pombe* MAM2 gene encodes a putative pheromone receptor which has a significant homology with the *Saccharomyces cerevisiae* STE2 protein. EMBO J. 1991;10(12):3743-3751.

50. Hagen DC, McCaffrey G, Sprague GF Jr. Evidence the yeast STE3 gene encodes a receptor for the peptide pheromone a factor: gene sequence and implications for the structure of the presumed receptor. Proc Natl Acad Sci U S A. 1986;83(5):1418-1422.

51. Tanaka K, Davey J, Imai Y, Yamamoto M. *Schizosaccharomyces pombe* map3+ encodes the putative M-factor receptor. Mol Cell Biol. 1993;13(1):80-88.

52. Sreedharan SP, Patel DR, Huang JX, Goetzl EJ. Cloning and functional expression of a human neuroendocrine vasoactive intestinal peptide receptor. Biochem Biophys Res Commun. 1993;193(2):546-553.

53. Dulac C, Axel R. A novel family of genes encoding putative pheromone receptors in mammals. Cell. 1995; 83(2):195-206.

54. Feldmann H, Aigle M, Aljinovic G, et al. Complete DNA sequence of yeast chromosome II. EMBO J. 1994;13(24):5795-5809.

55. Aljinovic G, Pohl TM. Sequence and analysis of 24 kb on chromosome II of *Saccharomyces cerevisiae*. Yeast. 1995;11(5):475-479.

56. Bieszke JA, Braun EL, Bean LE, Kang S, Natvig DO, Borkovich KA. The nop-1 gene of *Neurospora crassa* encodes a seven transmembrane helix retinal-binding protein homologous to archaeal rhodopsins. Proc Natl Acad Sci U S A. 1999;96(14):8034-8039.

57. Tusn‡dy GE, Simon I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol. 1998;283(2):489-506.

58. Brenner SE, Chothia C, Hubbard TJP. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A. 1998;95:6073-6078.

59. Retief JD. Phylogenetic analysis using PHYLIP. Methods Mol Biol. 2000;132:243-258.

60. Felder CB, Graul RC, Lee AY, Merkle H-P, SadŽe W. The venus flytrap of periplasmic binding proteins: an ancient protein module present in multiple drug receptors. PharmSci. 1999;1(2):http://www.pharmsci.org/journal/.

61. Engelke G, Gutowski-Eckel Z, Hammelmann M, Entian KD. Biosynthesis of the lantibiotic nisin: genomic organization and membrane localization of the NisB protein. Appl Environ Microbiol. 1992;58(11):3730-3743.

62. Park J, Karplus K, Barrett C, et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol. 1998;284:1201-1210.

63. Bauer H, Mayer H, Marchler-Bauer A, Salzer U, Prohaska R. Characterization of p40/GPR69A as a peripheral membrane protein related to the lantibiotic synthetase component C. Biochem Biophys Res Commun. 2000;275(1):69-74.